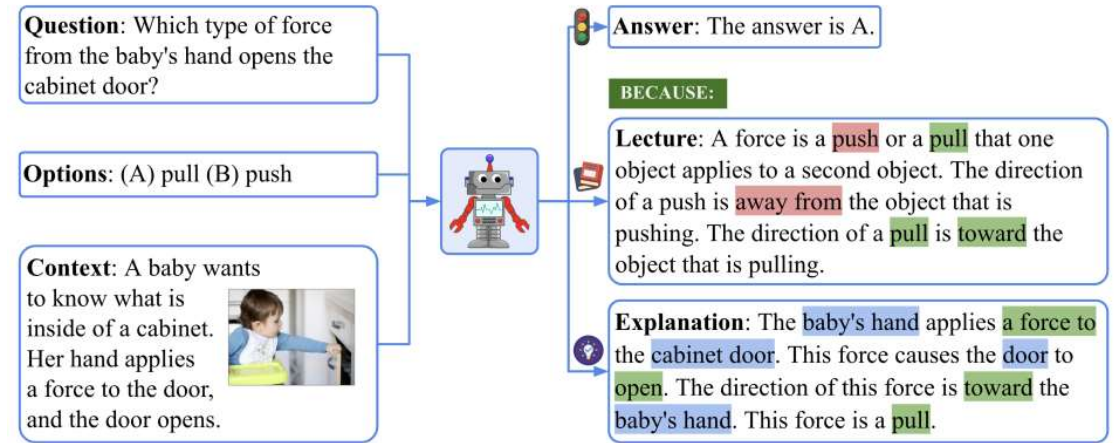


# Low Resource Offline Multimodal Chatbot for ScienceQA



Team BOTZZ ☐

David Liu  
Jazmin Guo  
Agrima Bansal  
Shihao Xiong  
Vansh Pruthi

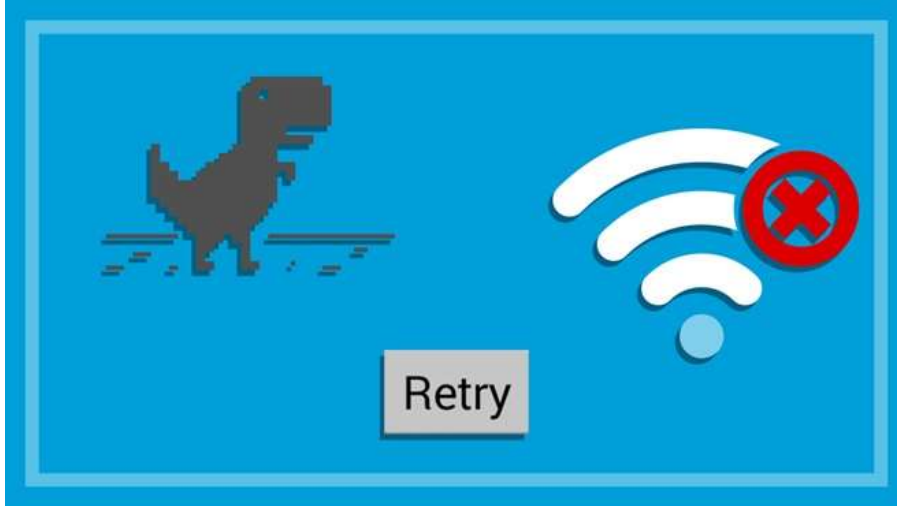
# Motivation and Problem Statement

1) Problem 2) Significance 3) Gap 4) Objective

# Motivation (1. problem)

Reliance on internet-based resources is not always feasible or desirable due to:

- Connectivity issues
- Privacy and safety concerns
- Distractions



# Motivation (2. significance)

Providing a free, accessible (low resource) and offline science learning tool can improve **safety and educational equity**.

1 December 2020 – Two thirds of the world's school-age children – or **1.3 billion children aged 3 to 17 years old** – do not have internet connection in their homes, according to a new joint report from UNICEF and the International Telecommunication Union (ITU).



Unicef UK

<https://www.unicef.org.uk> › press-releases › two-thirds-... ⋮

**Two thirds of the world's school-age children have no internet ...**

# Motivation (3. Gap)

There is a lack of specialised science chatbot that run efficiently and entirely offline on personal devices like laptops or mobile phones. While online proprietary state-of-the-art chatbots exist (such as ChatGPT), they rely on:

- Billed APIs
- Data collections
- Constant internet connectivity.

The screenshot displays the Claude AI pricing interface. On the left, two plans are compared: 'Plus' at \$20 USD/month and 'Pro' at \$200 USD/month. The 'Pro' plan is highlighted with a 'Get Pro' button. In the center, the 'Gemini Advanced' offer is shown at \$19.99 USD/month. On the right, the 'Subscribe to annual plan' section shows a discount from USD 216 to USD 180 per year. Below this, a list of reasons to switch to the Pro annual plan is provided.

Plus	Pro
<b>\$20</b> USD/month	<b>\$200</b> USD/month
Level up productivity and creativity with expanded access	Get the best of OpenAI with the highest level of access
<a href="#">Your current plan</a>	<a href="#">Get Pro</a>
<ul style="list-style-type: none"><li>✓ Everything in Free</li><li>✓ Extended limits on messaging, file uploads, advanced data analysis, and image generation</li><li>✓ Standard and advanced voice mode</li><li>✓ Limited access to o1 and o1-mini</li><li>✓ Opportunities to test new features</li><li>✓ Create and use custom GPTs</li><li>✓ Limited access to Sora video generation</li></ul>	<ul style="list-style-type: none"><li>✓ Everything in Plus</li><li>✓ Unlimited access to o1, o1-mini, and GPT-4o</li><li>✓ Unlimited access to advanced voice</li><li>✓ Access to o1 pro mode, which uses more compute for the best answers to the hardest questions</li><li>✓ Extended access to Sora video generation</li></ul>
<a href="#">Manage my subscription</a> <a href="#">I need help with a billing issue</a>	<a href="#">I need help with a billing issue</a> Usage must be reasonable and comply with our <a href="#">policies</a>

**Gemini Advanced**  
**\$19.99** USD / month

- ✓ With Ultra 1.0 model, our most capable AI model
- ✓ State-of-the-art performance
- ✓ Designed for highly complex tasks
- ✓ Available soon: Gemini in Gmail, Docs, and more

**Subscribe to annual plan**

[Offer expires in 2 days](#)

**USD 216 USD 180** /year + tax  
~~USD 18~~ USD 15 /month + tax

**Why switch to Pro annual now?**

- ✓ Save USD 60 your first year compared to monthly plan
- ✓ Lock in current pricing for a full year
- ✓ More usage than Free
- ✓ Organize documents and chats with Projects
- ✓ Access additional Claude models
- ✓ Use Claude 3.7 Sonnet with extended thinking mode

# Motivation (4. objective)

To develop a chatbot utilising a low resource open source LLM model (Gemma-3-4B-instruction-tuned) that **operates completely offline on user devices**. Low computational cost

- High accuracy



# Literature Review



# Literature Review

Introduce **ScienceQA**: A large-scale dataset with questions from science curricula

When a question is enriched with **Chain-of-Thought (CoT)** reasoning we see an improvement in accuracy

Model and prompt type	Accuracy
UnifiedQA with zero-shot	70.12%
GPT-3 with zero-shot	74.04%
UnifiedQA with CoT	74.11 (3.99↑)
GPT-3with CoT	75.17 (1.20↑)

Our project uses **ScienceQA** dataset and **CoT** prompts.

---

## Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering

---

Pan Lu<sup>1,3</sup>, Swaroop Mishra<sup>2,3</sup>, Tony Xia<sup>1</sup>, Liang Qiu<sup>1</sup>, Kai-Wei Chang<sup>1</sup>,  
Song-Chun Zhu<sup>1</sup>, Oyvind Tafjord<sup>3</sup>, Peter Clark<sup>3</sup>, Ashwin Kalyan<sup>3</sup>  
<sup>1</sup>University of California, Los Angeles, <sup>2</sup>Arizona State University, <sup>3</sup>Allen Institute for AI  
{lupantech, kwchang.cs}@gmail.com, sczhu@stat.ucla.edu,  
{oyvindt, peterc, ashwinkv}@allenai.org

Abstract



# Literature Review

Introduce Retrieval-Augmented Generation (RAG):

A hybrid architecture that combines the strengths of **parametric memory** with **non-parametric memory**

**This helps to:**

- Improve factual accuracy and adaptability
- Generate more compatible and flexible information

We are using **retrieval (ChromaDB)** to bring in relevant lecture notes

---

## Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

---

Patrick Lewis<sup>†‡</sup>, Ethan Perez<sup>\*</sup>,

Aleksandra Piktus<sup>†</sup>, Fabio Petroni<sup>†</sup>, Vladimir Karpukhin<sup>†</sup>, Naman Goyal<sup>†</sup>, Heinrich Küttler<sup>†</sup>,

Mike Lewis<sup>†</sup>, Wen-tau Yih<sup>†</sup>, Tim Rocktäschel<sup>†‡</sup>, Sebastian Riedel<sup>†‡</sup>, Douwe Kiela<sup>†</sup>

<sup>†</sup>Facebook AI Research; <sup>‡</sup>University College London; <sup>\*</sup>New York University;  
plewis@fb.com

**Abstract**

# Literature Review

- Introduce a prompting technique: **Fact-and-Reflection (FaR)**
- The technique split the process into
  - 1. **Fact generation: Model recall the relevant knowledge**
  - 2. **Reflection: Model reasons base on the facts**

This help to:

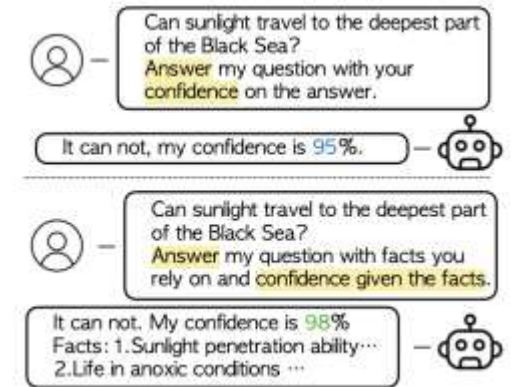
1. Avoid overconfidence and express uncertainty
2. Improve confidence that match the actual correctness
3. Get more trustworthy confidence estimates

## Fact-and-Reflection (FaR) Improves Confidence Calibration of Large Language Models

Xinran Zhao<sup>1,2,\*</sup> Hongming Zhang<sup>1</sup> Xiaoman Pan<sup>1</sup> Wenlin Yao<sup>1</sup>  
Dong Yu<sup>1</sup> Tongshuang Wu<sup>2</sup> Jianshu Chen<sup>1</sup>  
<sup>1</sup>Tencent AI Lab, Bellevue, <sup>2</sup>Carnegie Mellon University

### Abstract

For a LLM to be trustworthy, its confidence level should be *well calibrated* with its actual performance. While it is now common sense that LLM performances are greatly impacted by prompts, the confidence calibration in prompting LLMs has yet to be thoroughly explored. In this paper, we explore how different prompting strategies influence LLM confidence calibration and how it could be improved. We conduct extensive experiments on six prompting methods in the question-answering context and we



# The ScienceQA Dataset

# Dataset

## ➤ ScienceQA Dataset

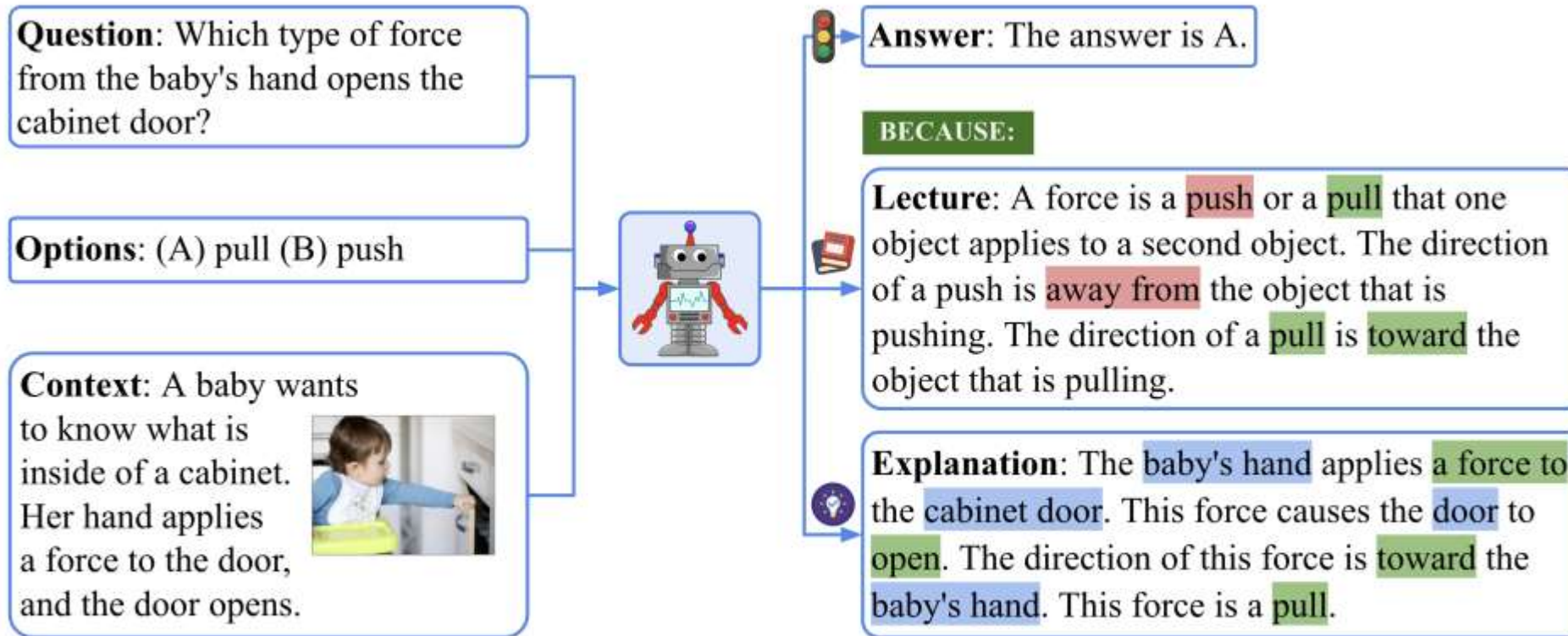
- The top image shows all the different fields in the dataset.
- Bottom image shows all the different topics and categories that the questions in the dataset cover
- This level of variety makes the ScienceQA dataset especially valuable for building general-purpose educational AI systems, or for specialized models focusing on specific school subjects.



<b>Biology</b> Genes to traits Classification Adaptations Traits and heredity Ecosystems Classification Scientific names Heredity Ecological interactions Cells Plants Animals Plant reproduction	<b>Physics</b> Materials Magnets Velocity and forces Force and motion Particle motion and energy Heat and thermal energy States of matter Kinetic and potential energy Mixture	<b>Geography</b> State capitals Geography Maps Oceania: geography Physical Geography The Americas: geography Oceans and continents Cities States	<b>History</b> Colonial America English colonies in North America The American Revolution <b>World History</b> Greece Ancient Mesopotamia World religions American history Medieval Asia	<b>Civics</b> Social skills Government The Constitution <b>Economics</b> Basic economic principles Supply and demand Banking and finance <b>Global Studies</b> Society and environment
<b>Earth Science</b> Weather and climate Rocks and minerals Astronomy Fossils Earth events Plate tectonics	<b>Chemistry</b> Solutions Physical and chemical change Atoms and molecules Chemical reactions <b>Engineering</b> Designing experiments Engineering practices <b>Units and Measurement</b> Weather and climate	<b>Writing Strategies</b> Supporting arguments Sentences, fragments, and run-ons Word usage and nuance Creative techniques Audience, purpose, and tone Pronouns and antecedents Persuasive strategies Editing and revising Visual elements Opinion writing	<b>Vocabulary</b> Categories Shades of meaning Comprehension strategies Context clues <b>Grammar</b> Sentences and fragments Phrases and clauses <b>Figurative Language</b> Literary devices	<b>Verbs</b> Verb tense <b>Capitalization</b> Formatting <b>Punctuation</b> Fragments <b>Phonology</b> Rhyming <b>Reference</b> Research skills

# Dataset

- Basic example of how the fields/metadata in the dataset are used





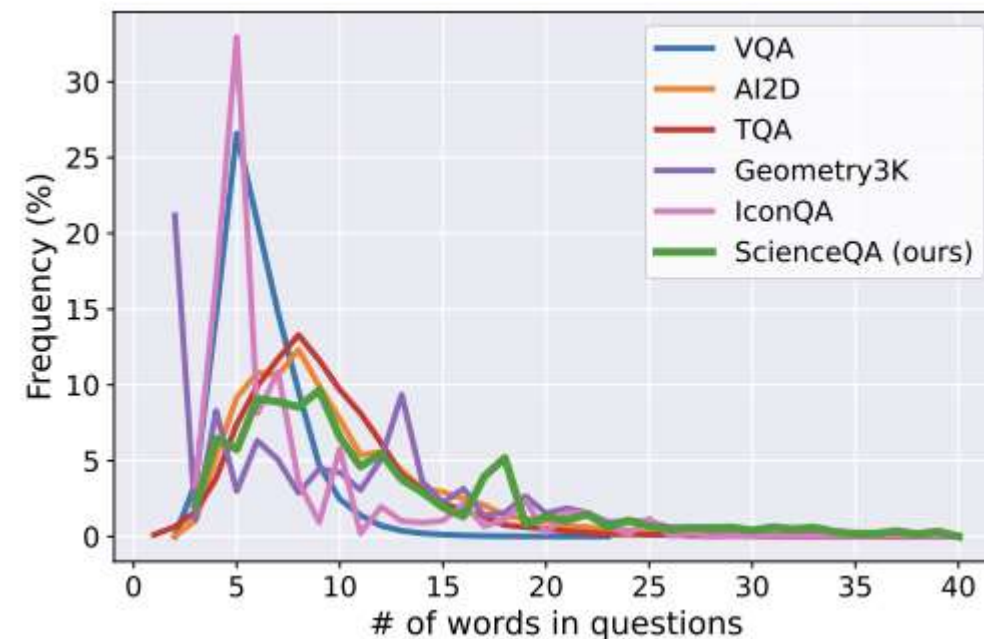
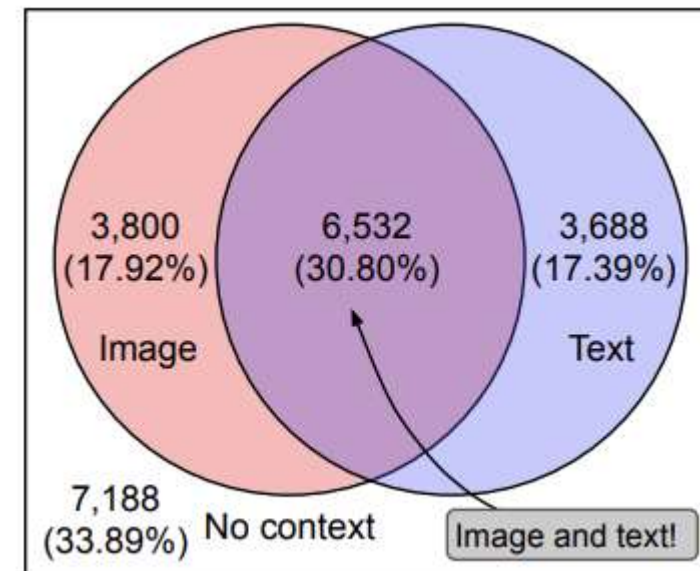
# Dataset

## ➤ Data Set Split

Split (3)

- ✓ train (12.7k rows)
- validation (4.24k rows)
- test (4.24k rows)

- The dataset has **~21K questions** that are randomly split into training, validation, and test splits with a ratio of **60:20:20**.
- In ScienceQA, (48.7%) have an image context, (48.2%) have a text context, and (30.8%) have both.
- Sometimes, the model is given a lot of information from multiple sources, while at other times, the only source of information is the question itself.



# Methods



# Methods

## 1. Experimental Setup & Rationale

- **Objective:** Improve the performance (accuracy and efficiency) of the model with different methods.
- **Dataset:** ScienceQA
- **Platform:** Google Colab

## 2. Main Methods

- Fine-tuning
- RAG(Retrieval Augmented Generation)

# Fine-tuning

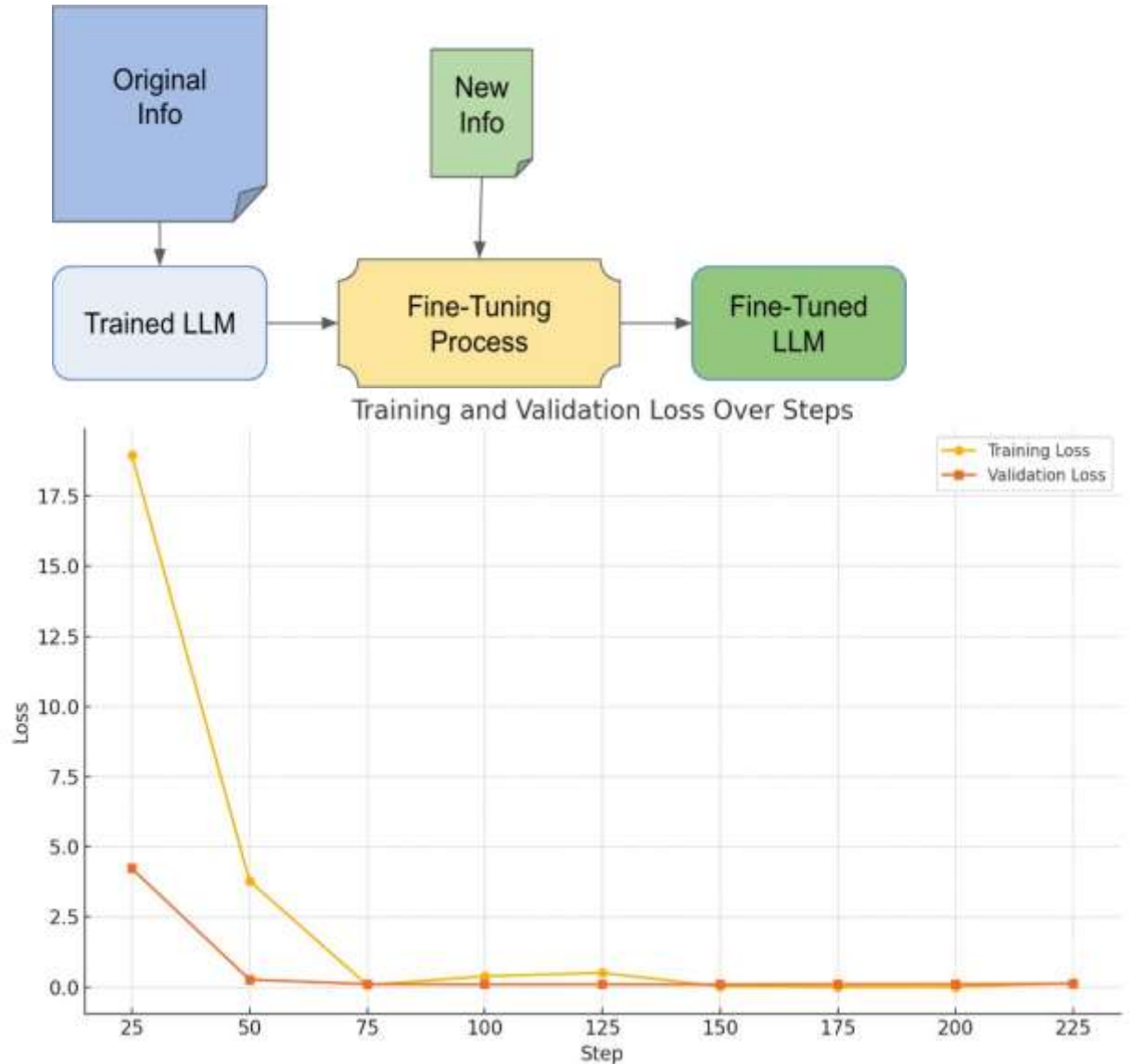
- Use the training and validation set of the data
- Format the data with labels (supervised)
- **PEFT** (parameter-efficient fine-tuning) library by Hugging Face
- LoRA (low-rank adaption)

## Difficulties:

- Passing text and image (multimodal input) to be tokenised and trained

## Drawback:

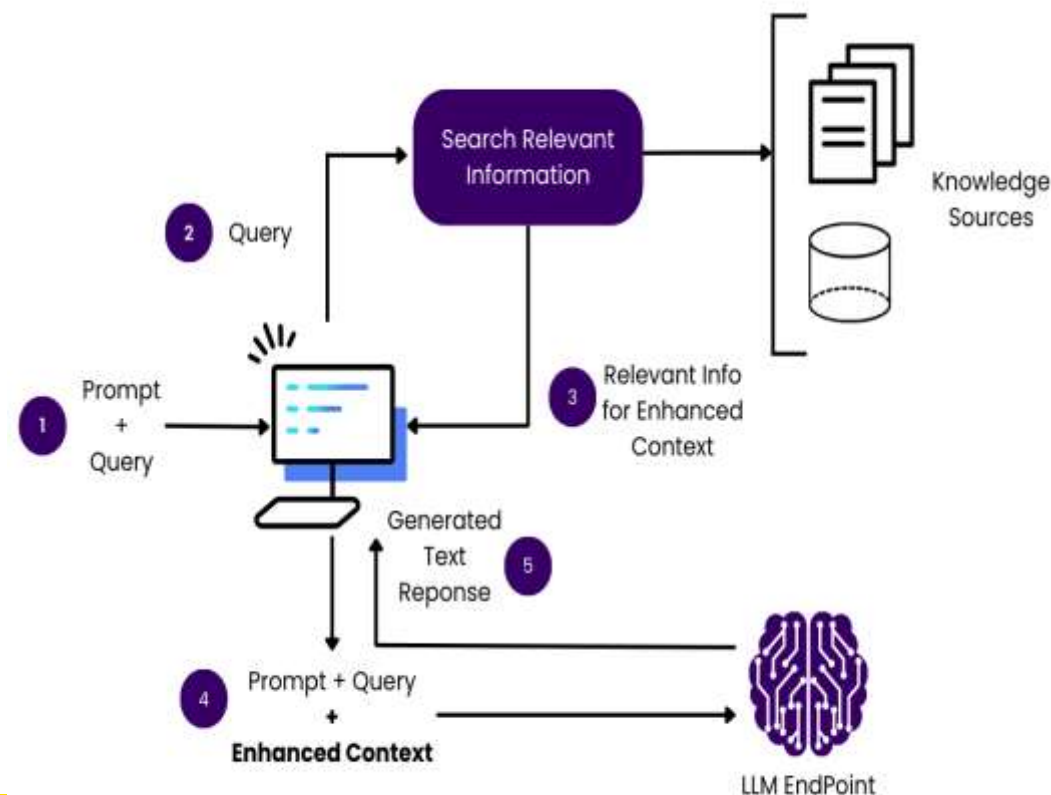
- Resource intensive -> **not ideal** for minor information updates



# RAG (Retrieval Augmented Generation)

**Definition:** RAG a technique that enhances the accuracy and relevance of Large Language Models by incorporating external knowledge sources

1. User query
2. Retriever (via vector DB like ChromaDB)
  - Converts into embeddings
  - Finds top-k relevant documents based on similarity
3. Augmented prompt
4. Retrieved documents+ original question are passed to LLM
5. Generator:
  - The model (e.g. GPT4o mini) generated a context aware answer.

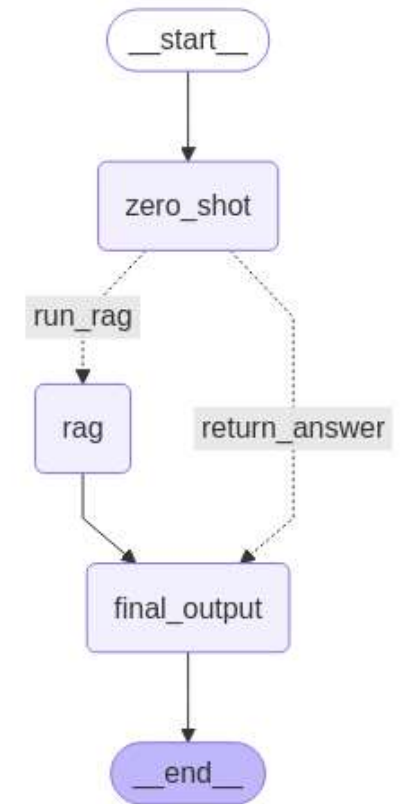


# RAG (Retrieval Augmented Generation)

- Sentence-transformers model
  - all-MiniLM-L12-v2
- maps sentences & paragraphs to a 384 dimensional dense vector space and can be used for tasks like clustering or semantic search.
- Vector DB
  - ChromaDB
  - Stores the vectors
  - Retrieve based on vector similarity and metadata

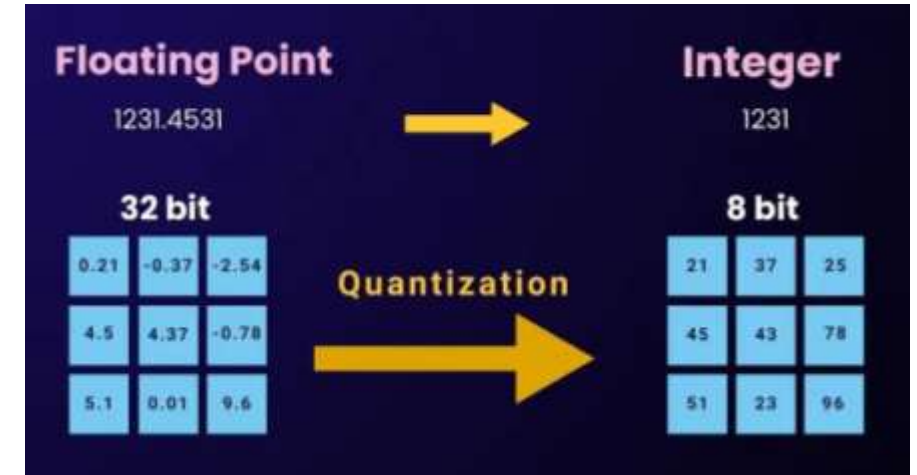
# Integration of methods with Langgraph

- Use model's zero shot confidence score to determine if langgraph is need
- High confidence -> return zero shot response
- Low confidence -> run RAG



# Quantization

- Load the model with 8-bit integer
- Quantization is a compression technique that involves mapping high precision values to a lower precision one
- Improves efficiency on consumer-grade hardware



# Results



# Using a subset of the data

1. After initial testing with 4o-mini (minimal prompting) we extracted a **subset of the training dataset** which contained 490 questions that 4o-mini got wrong. This is **12%** of the training set.
2. We used this subset (the "difficult" questions) to test various methods.



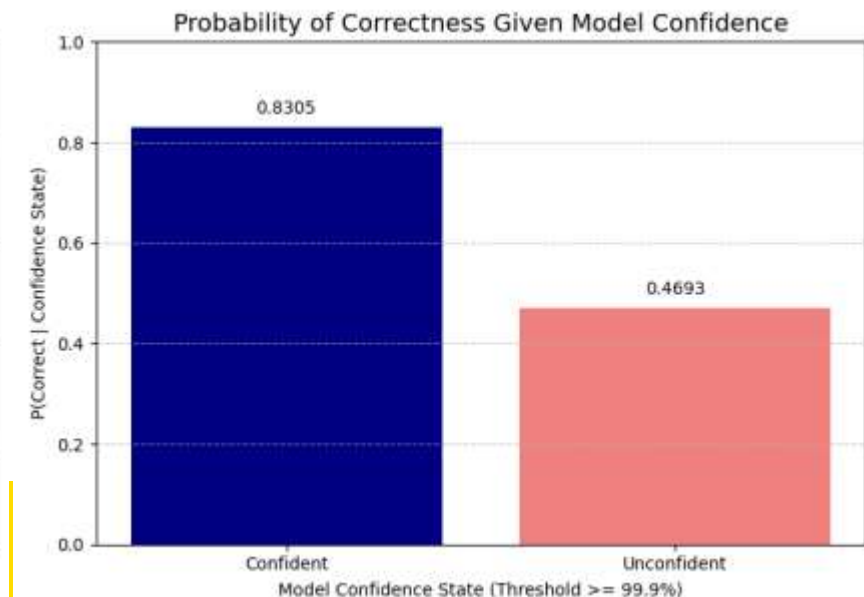
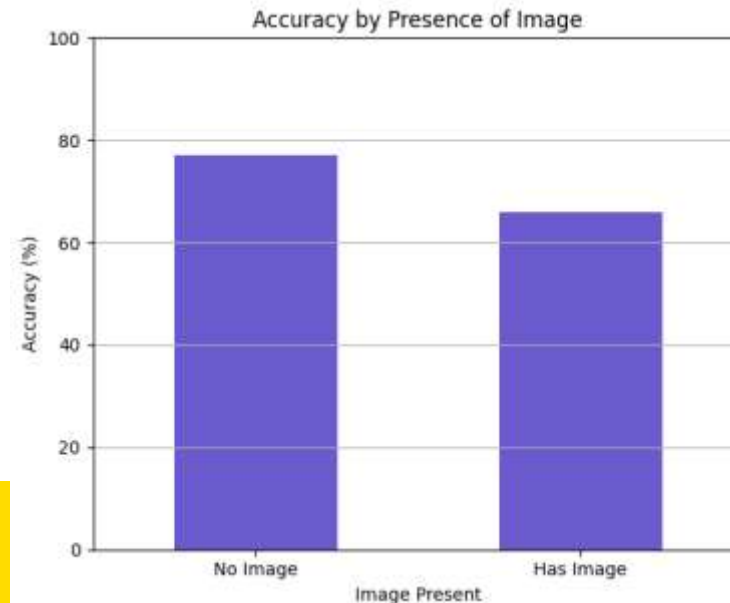
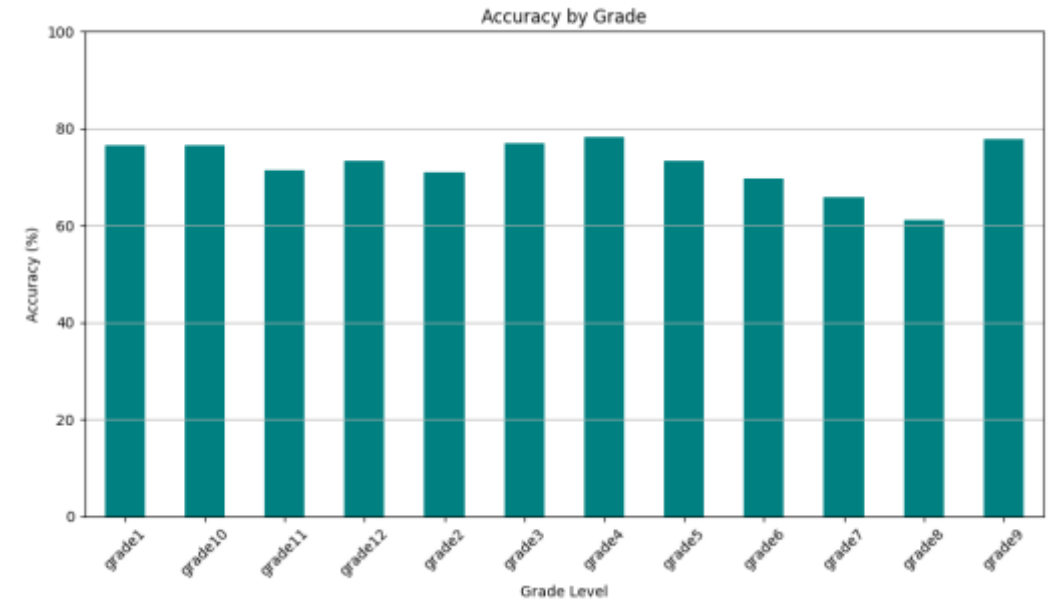
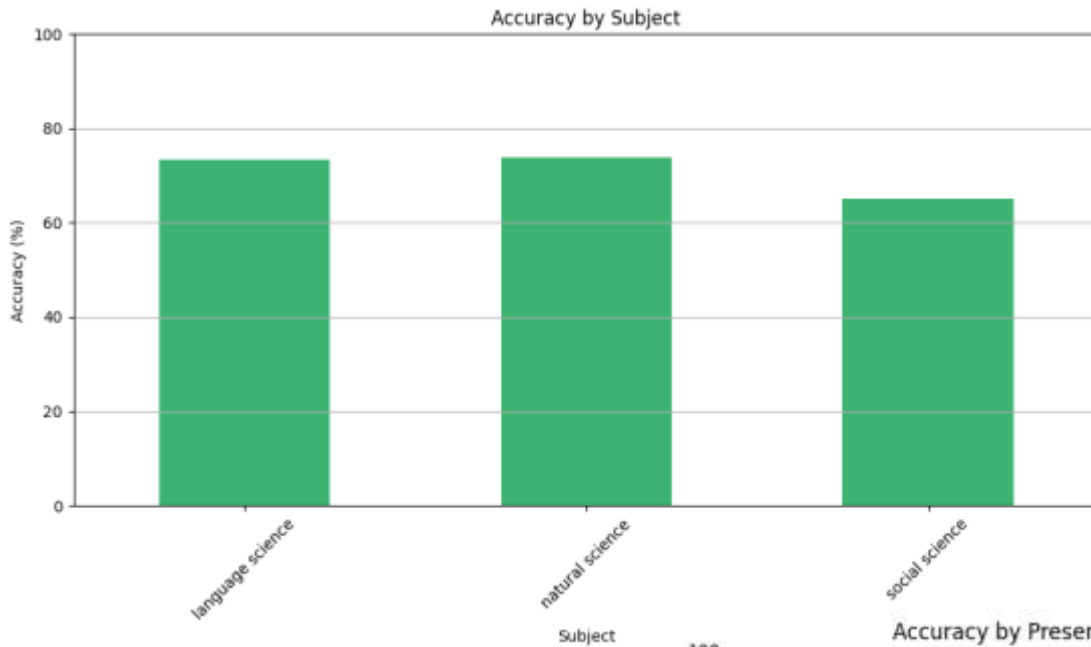
# Results

(no fine tuning,  
zero-shot vs  
RAG)

Model Name	Accuracy (on “difficult” questions subset size=490)
GPT-4o	49.28%
Gemini-2.0-Flash	55.21%
Gemini-2.0-Flash (with RAG #1)	60.82%
Gemini-2.0-Flash (with RAG #2)	57.96%
Gemini-2.0-Flash-Lite	56.03%
Gemini-2.0-Flash-Lite (with RAG #1)	58.75%
Gemini-2.0-Flash-Lite (with RAG #2)	58.78%

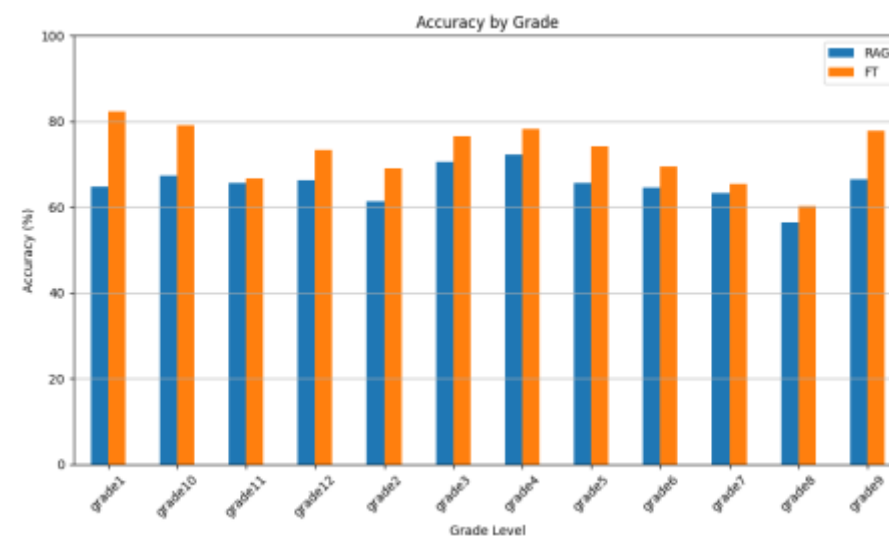
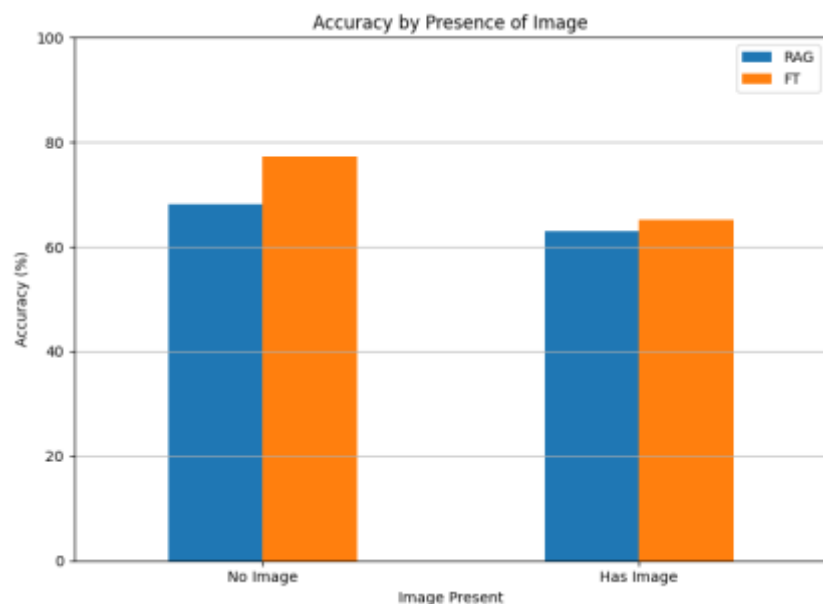
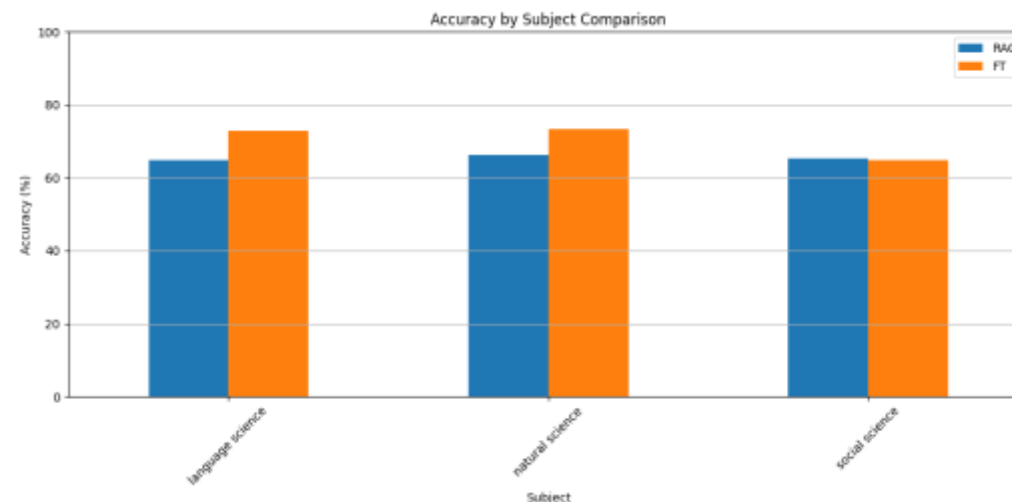
RAG method #	K (retrievals)	Embedding Method	VectorDB Size
1	1	text-embedding-3-large	12726
2	1	sentence-transformers/all-MiniLM-L12-v2	234 (deleted duplicates)

# Data Analysis (Using gemma-3-4b it with minimal prompting)



# Results – RAG vs Fine tuned

Model Name	Accuracy
RAG	65.64%
Fine Tuned	71.52%



# Data Analysis

- Found a correlation between confidence and accuracy.
- When  $>99.9\%$  confident, accuracy is Base : 83%, FT : 84%
- When  $<99.9\%$  confident, accuracy is Base : 47%, FT : 47%

```
P(correct | confident): 0.83  
P(correct | unconfident): 0.47
```

Gemma 4B it Base model  
Confident threshold: confidence $>0.999$

```
P(correct | confident): 0.84  
P(correct | unconfident): 0.47
```

Gemma 4B it Fine Tuned  
Confident threshold: confidence $>0.999$

# RAM usage quantized vs Full Size

Python 3 Google Compute Engine backend (GPU)  
Showing resources from 9:12 PM to 9:24 PM

System RAM  
6.1 / 53.0 GB



GPU RAM  
6.6 / 22.5 GB



Disk  
50.4 / 235.7 GB



Gemma 4B quantized

Python 3 Google Compute Engine backend (GPU)  
Showing resources from 9:17 PM to 9:24 PM

System RAM  
3.9 / 53.0 GB



GPU RAM  
18.9 / 22.5 GB



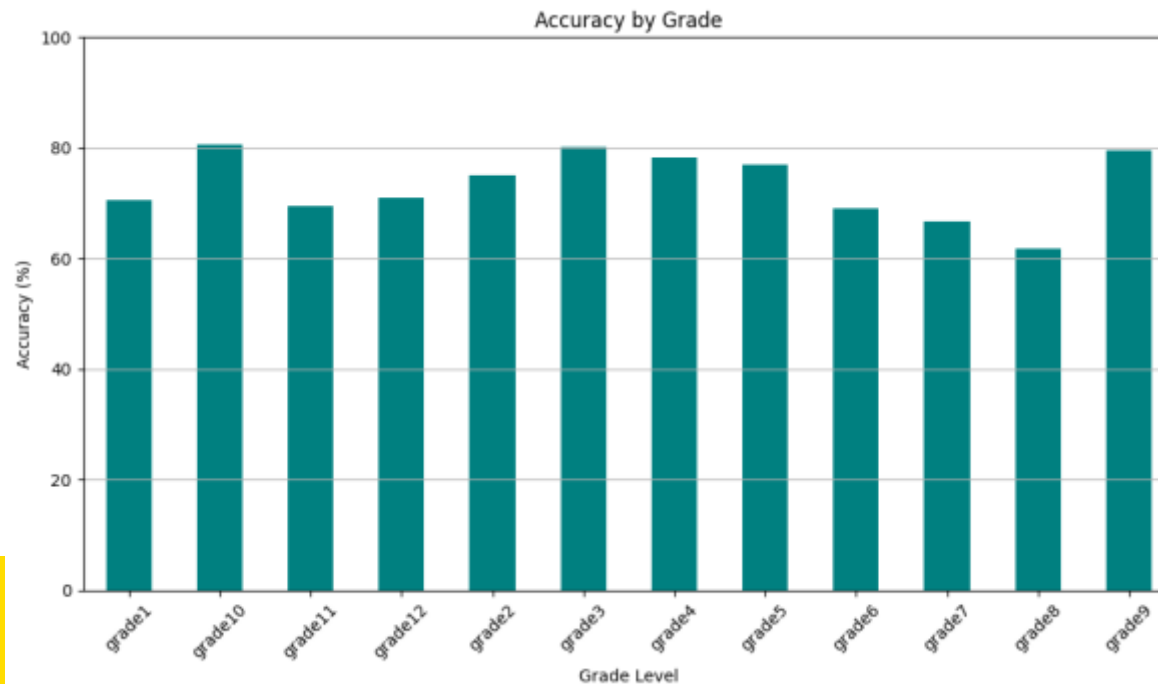
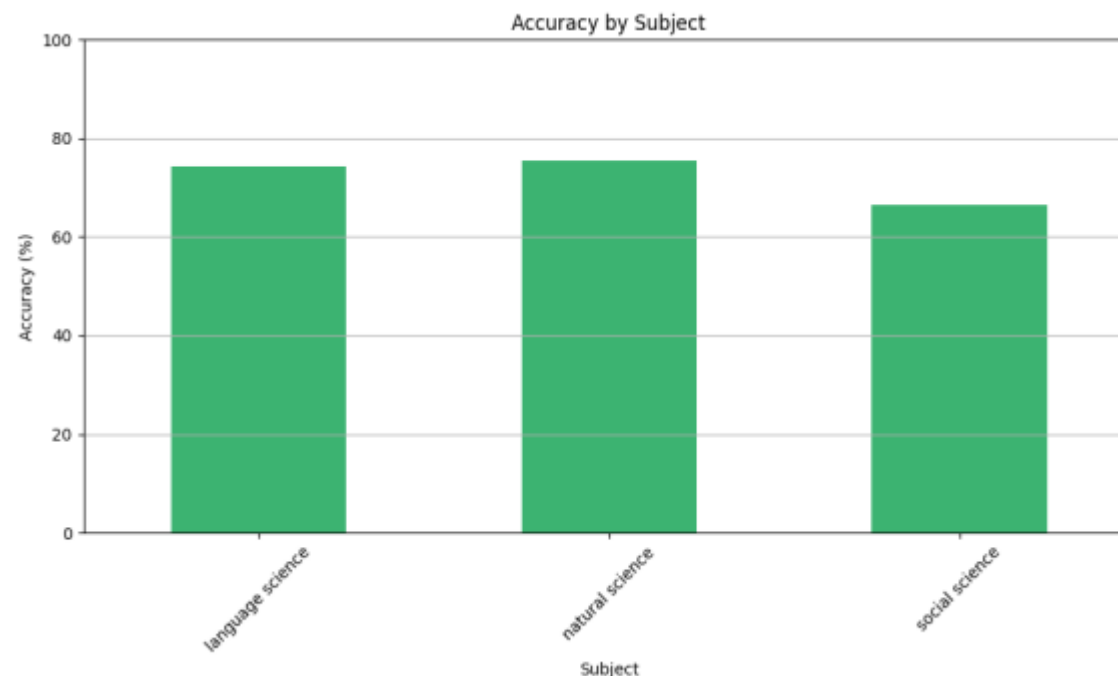
Disk  
50.1 / 235.7 GB



Gemma 4B full size

# Mixed Method

Model Name	Accuracy
Mixed Quantised	73.31%





Gemma 3 Variant	Accuracy on test set
Fine tuned mixed methods quantised	73.31% (0.10 ↑)
Base model zero shot quantized	73.21%

# Discussion and Conclusion

# Discussion and Conclusion

- Gemma 3-4b performed significantly worse with RAG
  - Small models like Gemma 3–4B can't handle long retrieved documents effectively
  - Small models are more affected by minor changes in how prompts and context are phrased
  - Quality of the prompt can significantly impact the output
- RAG is a contingency for when the model has low confidence
- Quantising the data reduces the RAM usage (makes it computationally less expensive)
- Our Fine Tuning did not improve the model's performance by much.